U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Substance Abuse and Mental Health Services Administration
Center for Substance Abuse Prevention
www.samhsa.gov

A GUIDE TO THE NREP REVIEW PROCESS

To assist its practice and policy-making constituents in learning more about evidence-based programs, SAMHSA's Center for Substance Abuse Prevention (CSAP) created a National Registry of Effective Programs (NREP). The purpose of NREP (modelprograms@samhsa.gov) is to create a repository of effective, evidence-based programs through rigorous scientific reviews of program evaluations, methodology, and findings. In 2002, SAMHSA's Model Programs Dissemination Project identified its first evidence-based programs designed to eliminate or reduce substance use and abuse in work settings.

To bring the best programs to the attention of the practice community, NREP continues to invite interventions, approaches, and curricula that address substance use and abuse in workplace settings. Those efforts may take many forms, such as employee assistance programs (EAPs), health/wellness/safety programs, drug-free workplaces, referral services, and programs to prevent and treat not only substance use but also interpersonal, traumatic, and family health/wellness and mental health issues associated with substance use.

NREP REVIEW CRITERIA FOR WORKPLACE PROGRAMS

NREP reviewers rate programs on a predetermined set of review criteria, each with its own 5-point rating scale.

Conceptual/Logic Model

The conceptual/logic model criterion is the degree to which the project findings are based in a clear and well-articulated model, either conceptual or logic-based. The possible ratings are as follows:

1 = No information about model

2 = Very little information about model

3 = Adequate information about model

4 = Nearly complete information about model

5 = Full and complete information about model

A surprisingly high number of manuscripts containing evaluations of prevention programs are published with little if any theoretical grounding for the implementations tested, other than an implicit appeal to common sense. The theoretical (or "conceptual") basis for an intervention provides an explanation of why and how it is expected to achieve its intended results and should be supported by prior conceptual development and research. An advantage

of theory-based interventions—in addition to providing a theoretical justification for the intervention to be tested—is that they suggest the various mechanisms by which the intervention is expected to affect its ultimate desired outcomes. That is, they specify the causal path of intervening mediators and moderators, or risk and protective factors. Logic models, which can be surprisingly challenging to articulate, are invaluable guides both to what constructs should be measured and to the analytic strategies by which program effects can be determined. If a program fails to achieve its intended outcomes, a good logic model can assist in examining which program components may have been at least partially effective in changing the intermediate objectives they target. The careful assessment of moderators may also help determine for what subpopulations a given intervention was most or least effective.

Intervention Fidelity

Intervention fidelity may include dosage data and evidence of adherence to program. The possible ratings are as follows:

0 = Non-applicable

1 = No or very weak evidence of fidelity to program

2 = Weak evidence of fidelity to program

3 = Some evidence of fidelity to program

4 = Strong evidence of fidelity to program

5 = Very strong evidence of fidelity to program

One of the inadequacies often identified by NREP reviewers of workplace-based programs is their failure to record carefully the fidelity with which the intervention to be studied was implemented. Indeed, the reviewer is often asked to take it on faith that the intervention was implemented in the manner in which it was designed. Even less often do evaluators assess other contemporaneous events—either isolated or ongoing—that might offer an alternate explanation of study results. Evaluators are becoming increasingly sophisticated about the importance of studying fidelity, both because complete fidelity is rarely achieved and because the empirical literature overwhelmingly links fidelity of implementation with desired outcomes. Evaluators also recognize that the concept of fidelity comprises multiple components, including the elements or activities of the intervention that are implemented, any adaptations or additions made, the relationship of the program administrator with study respondents, the frequency and length of time over which the intervention was implemented, and the individual dose that each respondent received. Failure to implement interventions as

planned is thought to be a prime cause of failure to find effects. However, even when effects are demonstrated, the absence of documentation of fidelity may leave questions concerning whether the intervention can be successfully replicated, because the adaptations made to it for the particular population to which it was targeted may be responsible for its success. The measurement of fidelity is no less important when the intervention to be implemented is a policy (e.g., relating to for-cause drug testing) than when it is a program; policies may (or may not) be administered uniformly, and consequences for infractions may be differentially applied.

The measurement of fidelity is itself now becoming a science. The simplest way in which to do so is to ask program administrators to report their activities on a regular basis, including what they are doing instead of or in addition to the program as intended. But such self-reports can be quite self-serving, and evidence is mounting that administrators who lack a full understanding of the program they are implementing may be unaware of the nature and extent of their modifications to it. Hence, a more optimal way to assess fidelity is through observation by unbiased observers.

Design

Design is the extent to which the research design was suitable for testing outcomes. The possible ratings are as follows:

1	=	No control or comparison group
2	=	Inappropriate (nonequivalent) control or comparison group; no attempt at either true or quasi-experimental design; comparison group inappropriate for a test of hypothesized outcomes
3	=	Control group or comparison group matched on some pertinent variables; somewhat appropriate for testing outcomes
4	=	Control group or comparison group appropriately matched on most pertinent variables; appropriate for testing outcomes
5	=	Excellent control or comparison group, either matched with treatment group on all pertinent variables or randomly assigned

Randomized controlled studies continue to be considered the “gold standard” of evaluative research. In randomized clinical trials, individuals or group-level units of which they are a part (e.g., worksites) are randomly assigned to an intervention or control group. The great virtue of randomized trials is that they distribute by chance to treatment and control

conditions characteristics associated with individuals or groups that might otherwise confound (i.e., serve as alternate explanations for) any differences between treatment and control group that the evaluation may find. These extraneous characteristics may be both observed and unobserved. In worksite settings in which uncontrolled evaluations are implemented, for example, individuals or worksites may self-select to receive a given intervention because of greater receptivity or perceived need. Or a treatment site may undergo profound organizational changes once the intervention has been initiated—for example, acquiring a new CEO who is unfriendly to "special projects"—that would greatly decrease its chances for successful implementation. Random assignment procedures greatly reduce opportunities for these confounders to adversely affect study outcomes—as long as sufficient units (e.g., individuals or worksites) are included in the pool to be allocated.

Randomized trials are unfeasible in many worksite settings. They may be precluded by management and unions; or the resources required to implement a given intervention may be available at some sites but not others. Effectiveness studies that rely on comparisons among units that have not been *randomly* assigned are commonly referred to as "quasi-experimental." Quasi-experimental comparison groups may have self selected into the comparison group, such as employees choosing not to attend a health fair, or they may be nonrandomly assigned by researchers, such as worksites that do not receive an intervention because of logistical constraints faced by the researchers.

With quasi-experimental designs, it is important to tease out the critical variables within the characteristics of the intervention and comparison sites that might confound the interpretation of study results. Some of these characteristics may be difficult to discern and even more difficult to assess. Many of these characteristics are more easily measured (e.g., respondent age, gender, race/ethnicity, job category, salary, and length of time at the company). Units of assignment (again, individuals or worksites) can sometimes be successfully batched together on these characteristics, or "blocked," prior to purposive nonrandom assignment, thus assuring greater equivalency across groups. Or when these differences are noted once the units have been assigned, they can be statistically controlled for during the analysis phase of the study. Such techniques cannot control for unobserved or unmeasured characteristics, so some degree of doubt almost always remains about whether conclusions drawn (and the strength of these conclusions) are accurate.

Sample Size and Units of Assignment

The possible ratings for sample size and unit of assignment are as follows:

0 = Non-applicable

1 = Sample size unspecified

2 = Entirely insufficient sample size

3 = Marginally sufficient sample size

4 = Sample size entirely sufficient, but no power analysis

5 = Sample size entirely sufficient, and convincing power analysis provided

The issue of how many respondents and/or sites to enroll in a study is a critical component of a study's design. An optimal sample size is sufficient to find significant program effects if they actually exist, but not so large as to be wasteful of study resources. Typically, the calculation of sample size—called a "power analysis"—is based on the ultimate desired outcome of interest and is dependent on several factors. These include the magnitude of the expected effect of the intervention and the sensitivity to change of the measures used to detect it. Both of these factors can be quite difficult to estimate if the program is new, or if information about the sensitivity of the measures to be used is unavailable. The calculation of power becomes more challenging when the units that are assigned to treatment or comparison groups are not individuals but groups or worksites. Then the number of groups becomes a key determinant of statistical power, along with the anticipated degree of within-group similarity of individuals. Even very large numbers of individuals cannot usually overcome the statistical limitations of having only a few sites, when sites (rather than individuals) are the assigned units.

Attrition

Attrition is the evidence of sample quality based on information about the rate at which study participants drop out of the study. The possible ratings are as follows:

0 = Non-applicable

1 = No data on attrition or very high attrition (81–100%)

2 = High attrition (61 to 80%)

3 = Moderate attrition (41 to 60%)

4 = Low attrition (21 to 40%)

5 = Very low attrition (0 to 20%)

There are several different types of attrition, all of which can impact program evaluations. NREP reviewers are sensitive to these issues. Enrollment attrition describes loss of study respondents between recruitment and program implementation. Program attrition pertains to loss from the intervention itself, and includes both attendance and dropout. Study attrition refers to respondents who fail to fully complete research protocols. All types of attrition should be carefully tracked and, of course, minimized; any of them can introduce biases that may affect the interpretability of study results. However, some level of attrition does, inevitably, occur, and its likelihood of occurrence is in direct proportion to the researcher's control over the respondent sample. In many worksites, that level of control is quite modest, but sometimes can be increased through the judicious use of incentives for both program attendance and research protocol completion, and by minimizing the burdens imposed by each.

Analyses of Attrition Effects

This criterion rates the appropriateness of methods to analyze attrition. The possible ratings are as follows:

0 = Non-applicable

1 = No analysis or inadequate analysis of attrition

2 = Cursory analysis of attrition effects

3 = Adequate and appropriate analysis of attrition effects

4 = Several analytic methods employed; analyses relatively thorough, most questions about attrition bias answered

5 = State-of-the-art methods employed; questions regarding attrition bias answered and biases themselves adjusted

The question driving the analysis of all attrition effects is: how do respondents who are lost to attrition differ from those who remain? To the extent that these differences are found to be minimal (and nonsignificant), reviewers will have greater confidence in study results. Concerns about attrition can be mitigated to some extent if it can be shown that attrition rates and characteristics of dropouts are similar in both the intervention and control conditions. Typical attrition analyses comprise comparisons on sociodemographic characteristics, but more thorough analyses include comparisons on other key contextual, mediating, or moderating variables, and baseline measures of outcomes, that may be more highly associated with the ultimate outcomes of interest. These differences help refine the description of the population for which the intervention may be effective and to which results may be generalized. Analysis of program attrition effects may be helpful in determining whether there appears to be a "dose-response" effect—that is, whether the effectiveness of the program appears to increase with exposure to its components.

Methods to Correct Biases

This criterion measures the degree to which biases from nonequivalence, attrition, or missing data were corrected. The possible ratings are as follows:

0 = Non-applicable

1 = No attempt to correct biases

2 = Inadequate attempts to correct biases

3 = Attempts to adequately correct biases

4 = Adequately corrects biases

5 = State-of-the-art methods used to correct biases

Issues and problems with evaluations need to be fully disclosed and their potential effects acknowledged. Sometimes these biases can at least be partially overcome by means of statistical controls. For example, non-equivalencies between treatment and control groups—which can even arise following random assignment—can be mitigated by introducing the differentiating variables as covariates in analyses. This strategy can also be adopted to help deal with other types of problems, including those introduced by attrition. Re-weighting of the sample to compensate for biases provides an alternative strategy.

Outcome Measures: Substantive Relevance

This criterion rates the relevance of outcome measures in the context of target population, theory or conceptual framework, and intervention goals. The possible ratings are as follows:

1 = No or insufficient information

2 = Poor choice of measures

3 = Adequate choice of measures

4 = Relevant measures

5 = Highly relevant measures

What constitutes a worthy outcome to assess and by which to judge the success of a workplace program? Reviewers' opinions on this matter vary considerably. Some strongly prefer measures of substance abuse or other behavioral outcomes that are thought to be associated with abuse, such as results of drug tests, or absentee or injury rates. Others look more closely at the specific, stated goals of the prevention program itself, like reductions in workplace stress, increases in productivity, or referrals to EAPs. Outcomes pertinent to health, wellness, safety, and mental health may also be included. Some argue that not only are measures of substance use and abuse per se infeasible in many workplace environments but that employers in search of effective prevention programs are primarily interested in those outcomes that most directly affect their business (e.g. injury rates, absenteeism, productivity, and employee turnover). We advise that measures of substance abuse be included but only where both practical and appropriate for the intervention; it is most important that the outcomes assessed are consistent with the objectives of the program to be evaluated and the interests of potential employers.

Outcome Measures: Psychometric Properties

This criterion rates the reliability and validity of outcome measures. The possible ratings are as follows:

1 = No or insufficient information

2 = Low psychometric qualities

3 = Mixed quality

4 = Good psychometric qualities

5 = Excellent psychometric qualities

There are multiple types of reliability and validity, all of which have a bearing on the quality of the measures used. Many evaluators assess internal consistency or homogeneity; others may include test-retest stability or reliability across raters. Internal consistency only has meaning within the context of a scale. A scale comprises a set of items or questions that relate to a given construct and are expected to hang together in some way: that is, respondents who answer one item in a particular way should then answer the other items in a similar fashion. Although opinions differ somewhat as to what constitutes an acceptable threshold of reliability, it is usually thought that a coefficient alpha value of anything less than .70 indicates an unacceptably weak measure, one in which the "noise" introduced by the inadequacy of the items threatens to overwhelm the desired "signal" detecting the construct to be measured. However, many constructs relating to behaviors (as opposed to, for example, attitudes or beliefs) are assessed by only single items, and thus tests of internal consistency are irrelevant. In that case—and, indeed, generally speaking—it is best to use measures that have been used effectively elsewhere, have acceptable psychometric properties, and can thus be cited. CSAP's Core Measures Initiative (CMI) comprises a useful compendium of such measures.

Missing Data

This criterion rates the quality of data collection (i.e., amount of missing data). The possible ratings are as follows:

0 = Non-applicable

1 = High quantity of missing data

2 = Somewhat high quantity of missing data

3 = Average amount of missing data

4 = Some missing data

5 = No or almost no missing data

Missing data present more of a problem in some studies than others. Studies that rely on surveys typically will have little missing data, assuming that most respondents are motivated to complete questionnaires of reasonable length that they begin. Studies that rely on archival data are much more likely to find that some of those data are missing or have been incorrectly entered or stored and so are unusable. Missing data can pose a substantial problem for traditional multivariate analysis procedures, since one value missing in a particular variable can delete an entire observation.

Treatment of Missing Data

This criterion rates the degree to which missing data were analyzed. The possible ratings are as follows:

0 = Non-applicable

1 = No attempt to analyze missing data

2 = Inadequate attempts to analyze missing data

3 = Attempts to analyze missing data

4 = Adequately analyzed missing data

5 = State-of-the-art analysis of missing data

There are a variety of ways to treat missing data. The easiest (and safest) is simply to acknowledge it within the context of the evaluation and to assess and then discuss candidly any likely biases that may result from losing a set of respondents in a particular analysis. More sophisticated methods exist, however, for imputing values for missing data. These, however, are not often seen, at least in workplace NREP applications received to date.

Outcome Data Collection

This criterion rates the quality of procedures for collecting outcome data. The possible ratings are as follows:

1	=	Very biased manner of data collection with high demand characteristics; data collected in haphazard manner without any standardization
2	=	Somewhat biased manner of data collection with some demand characteristics; data collected in haphazard manner without any standardization
3	=	Relatively unbiased manner of data collection; standardized method of data collection; structured outcomes measures used specific to behaviors being investigated
4	=	Anonymous or confidentiality ensured in data collection; standardized method of data collection; structured, validated outcome measures used for assessing changes in target and non-target (discriminant) symptom areas
5	=	Anonymous or confidentiality ensured in data collection; standardized method of data collection; ethnic group or gender match between data collectors and participants specified; structured, validated outcome measures used for assessing changes in target and non-target (discriminant) symptom areas; measures used to examine for differential outcome according to subject characteristics

There are a variety of ways in which to collect data, which may introduce varying degrees of bias. Most desirable are totally objective mechanisms, especially on sensitive issues (like drug use), such as drug testing results or other objective assessments (e.g., obvious alcohol or drug impairment at work, archival record data such as DUI arrests). Less desirable are anonymous self-reports, because even though subjects may be assured that their responses can never be attributed, they may still find reasons to be less than fully disclosing. Least desirable are confidential self-reports, because respondents may not trust researchers' commitment to ensure privacy. Respondents' (and reviewers') concerns, however, can be somewhat allayed by careful descriptions of any measures taken to enhance respondents' beliefs concerning confidentiality, including adequate privacy, multiple assurances of confidentiality and the limited uses to which study data will be put, administration of protocols by staff who are unrelated to program implementation, and so on. While conditions under which data are collected typically are the same for both treatment and comparison groups, suspicion remains that those exposed to the intervention may respond to greater implicit pressure to report what they understand to be a desired outcome.

Analysis

The analysis criterion rates the appropriateness and technical adequacy of analytic techniques. The possible ratings are as follows:

1 = No analyses reported; all analyses inappropriate or do not account for important factors

2 = Some but not all analyses inappropriate or left out important factors

3 = Mixed in terms of appropriateness and technical adequacy

4 = Appropriate analyses, but not cutting edge techniques

5 = Proper state-of-the-art analyses conducted, included subgroup analyses

Analyses do not have to be sophisticated to be worthy. Indeed, there is some virtue to using the most parsimonious and easily explained analytic strategy, consistent with the question that is being addressed and the particular constraints inherent in the data. Key elements of the analysis are a description of the group differences observed in the intermediate and ultimate outcome measures, and a proper assessment of the statistical significances of those differences. Typically, multivariate analyses that control carefully for extraneous factors on which treatment and comparison groups differ, and address the various pathways and influences specified in the program's logic model, are entirely sufficient. However, reviewers will look carefully to ensure that appropriate techniques were used to ensure that analyses took into account the unit of assignment, if that was at a group level. Reviewers are also likely to determine if initial analyses encompassed all the individuals assigned to an intervention or comparison group (i.e., "intent-to-treat"), regardless of whether they subsequently failed to enroll in the intervention or dropped out of it. In addition, reviewers may look for subgroup analyses to discover if a modest main effect is disguising a particularly strong effect for a particular segment of the population studied (e.g., those at high risk of substance abuse) and little or no effects for other subgroups.

Outcomes

This criterion measures the degree to which findings support study hypotheses. The possible ratings are as follows:

1 = Findings contradict, or clearly do not support, pertinent study hypotheses

2 = Findings provide minimal evidence supporting pertinent study hypotheses but include other null or contradictory findings

3 = Pertinent findings are generally and consistently in the direction predicted by study hypotheses but do not reach statistical significance

4 = Findings for the most part reach statistical significance but are not robust or uniform across outcome domains

5 = Findings unequivocally and consistently support pertinent study hypotheses

This criterion addresses the simple question, "Did it work?" That is, did the intervention achieve its desired outcomes? Here, the astute reviewer will examine both what the evaluator does and does not report: that is, effects on some outcomes but not others, as well as the relative importance of the outcomes reported. The evaluator is thus strongly advised to report findings for every outcome measured, and the reviewer will then look for a consistent pattern of results, even when all do not reach statistical significance. The reviewer will also examine the magnitude of any statistically significant outcomes, especially in light of the size of the sample, for the larger the sample, the easier it is to detect such differences: thus, a finding may have statistical but not practical significance.

Other Plausible Threats to Validity

This criterion is the degree to which the design addresses and eliminates plausible alternative hypotheses concerning program effects; degree to which design warrants causal attributions. The possible ratings are as follows:

1 = Very high threat to validity; inability to attribute effects to program

2 = Substantial threat to validity; difficult to attribute effects to program

3 = Moderate of threat to validity; mixed ability to attribute effects to the program

4 = Low threat to validity; fairly high ability to attribute effects to program

5 = No or very low threat to validity; high ability to attribute effects to program

In this catch-all category, NREP reviewers consider all the residual issues and problems, not covered elsewhere, that might adversely affect their confidence to attribute findings to the intervention instead of other causes. For instance, given the rapidity with which outcomes decay over time, reviewers may assess the timing of study post-tests, giving higher credibility to those that are administered some period (e.g., 6 months) following the program's completion. Reviewers may also reiterate or reframe concerns noted previously.

Integrity

Integrity is the overall level of confidence in project findings based on research design and implementation. The possible ratings are as follows:

1 = No confidence in results

2 = Weak, little confidence in results

3 = Mixed, some weak, some strong characteristics

4 = Strong, fairly good confidence in results

5 = High confidence in results; findings fully defensible

This represents the first of two subjective, summary judgments that encompasses the overall quality of the methodology of the evaluation, and thus the reviewer's ability to attribute its findings to the intervention and not to some set of extraneous causes. It constitutes a consideration not only of all of the characteristics specified above but of any other issues that the reviewer may notice, such as the length of time between pre- and post-tests.

Utility

Utility is the overall usefulness of project findings to inform prevention theory and practice. Ratings are anchored according to the following categories and combine strength of findings and strength of evaluation. The possible ratings are as follows:

1 = Clear findings of null or negative effects for a program with well-articulated theory and well-implemented program design; study provides support for rejecting the program as a replication model

2 = Findings predominately null or negative, though not uniform or definitive

3 = Ambiguous findings because of inconsistent result or methods weaknesses that do not provide a strong basis for programmatic or theoretical contributions

4 = Positive findings that demonstrate the efficacy of the program in some areas, or support the efficacy of some components of the program

5 = Clear findings supporting the efficacy of well-articulated theory and program design, the study provides support for the program as a replication model

This represents the second of these two summary judgments and pertains to the overall strength and applicability of the evaluation's findings. Note that this criterion incorporates some of the methodological issues included in the previous one (integrity) but primarily addresses whether these findings are both positive and consistent across pertinent domains.

Replications

This criterion rates the number of adaptations of the model in different settings and/or by different workplaces, evaluators, etc., with similar positive results of both the intervention implementation and evaluation. The possible ratings are as follows:

1 = No replication; study reviewed represents program's only available evaluation

2 = One self-replication by program developer in different site with similar positive results; one replication but no independent evaluator

3 = Two or more self-replications by program developer in different sites with similar positive results

4 = One or two replications by independent evaluators in different sites with similar positive results

5 = Three or more replications by independent evaluators producing similar positive results

The independence of evaluations, and their replications, continues to be an issue within prevention evaluation. Many prevention programs classified as effective, model, or promising by NREP have been evaluated only once, either by their developers or by evaluators working in close collaboration with their developers. As such, the resulting evaluation cannot truly be called independent and is subject to at least the appearance of bias—that is, the limitation of findings published to those that demonstrate program success. Even more problematic, the findings reported from solitary evaluations may be limited to the particular population studied or may be a function of the attention and resources lavished on the administration of the program by the developer. Even when a program is implemented and evaluated a second time within the context of another population, it has often undergone substantial revisions, so the set of evaluations that is submitted for review pertain more to an approach than a program administered with a consistent set of protocols. Thus, very few applications are rated a "4" or "5" on this criterion, and it is not explicitly considered in the previous rating of utility at this time.

Dissemination Capability

This criterion rates the materials developed, including training in program implementation, technical assistance, standardized curriculum and evaluation materials, manuals, fidelity instrumentation, videos, recruitment forms, etc. The possible ratings are as follows:

0	=	Non-applicable
1	=	Materials, training, and technical assistance not available; in case of model that requires no curriculum (i.e., therapeutic models), training/qualified trainers and technical assistance not available
2	=	Materials available but of low quality or very limited in scope; training/qualified trainers and technical assistance either not available or limited
3	=	Materials of sufficient quality with limited technical assistance and/or training/qualified trainers
4	=	High quality materials, limited technical assistance and/or training/qualified trainers or vice versa
5	=	High quality materials, technical assistance and training/qualified trainers readily available

This criterion concerns the readiness of effective prevention programs for prime time (i.e., model status). Almost all effective programs require an infrastructure of support to ensure that practitioners understand how to implement them with fidelity, and where adaptations may

be made without compromising program effectiveness. Typically, this infrastructure includes the availability of initial training and ongoing technical assistance, as well as standardized manuals and protocols.

REVIEW PROCESS

After application materials are received by the NREP office, applicants are contacted to confirm receipt of the materials and to verify that they are complete. At this time, additional materials may be requested. The review begins with a triage process that culls those applications that are clearly inappropriate for review; these are returned to the applicant with encouragement to address the issues noted and reapply at some future time.

The remaining applications are submitted to *ad hoc* teams of three reviewers who have been trained in the NREP process and have demonstrated expertise in the field of workplace substance abuse prevention, early intervention, and treatment. All reviewers have terminal degrees in their respective fields, and most have received grants from NIH and have either academic appointments or work in private research settings.

Reviewers independently assess the materials submitted and rate them on each listed criterion. These reviews are then collected by a fourth reviewer, who looks them over and identifies any substantial disparities in ratings. If these are found, the reviewers then caucus by telephone to share and defend their respective ratings and bring the ratings into alignment. Decisions are made through a consensual process.

Individual scores from members of each reviewer team are then compiled into a single document, together with their narrative descriptions of the review program's strengths, weaknesses, and major outcomes. As a final step, summary scores from the two critical parameters of integrity and utility are used to rate programs respectively on the scientific rigor of their evaluation methodology and the strength and practicality of their findings.

Averaged scores across raters for these two rating criteria are then used to classify programs as lacking in sufficient current support, promising, or effective. Programs defined as effective have the further option of being recognized as model if their developers choose to take part in SAMHSA dissemination efforts. The review requirements for each category are:

- **Insufficient Current Support** refers to programs that require additional data or details before they can receive a score warranting a level of Effective or Promising on *either* the summary judgments of Integrity or Utility. These programs may be very worthwhile and have many implications to inform other prevention, treatment, or

rehabilitation efforts, but in their current form they do not have sufficient evidence to warrant a rating of Promising or higher.

- **Promising Programs** have been implemented and evaluated sufficiently and are considered to be scientifically defensible. They have demonstrated positive outcomes in preventing substance abuse and related behaviors. However, they have not yet been shown to have sufficient rigor and/or consistently positive outcomes required for Effective Program status. Nonetheless, Promising Programs are eligible to be elevated to Effective or Model status subsequent to review of additional documentation regarding program effectiveness. Promising Programs must score at least 3.33 on *each* of the parameters of Integrity and Utility.

- **Effective Programs** are well-implemented, well-evaluated programs that produce a consistently positive pattern of results. Developers of Effective Programs have yet to agree to work with SAMHSA/CSAP to support broad-based dissemination of their programs but may disseminate their programs themselves. These programs must score at least 4.0 on a 5-point scale on *each* of the parameters of Integrity and Utility.

- **Model Programs** are effective programs whose developers have coordinated and agreed with SAMHSA to provide quality materials, training, and technical assistance for nationwide implementation. That help is essential to ensure that the program is carefully implemented, and maximizes the probability of repeated effectiveness.

Once a decision has been reached, the lead reviewer typically compiles all narrative comments pertaining to each criterion, and the mean score for that criterion, into a document that is shared with (and only with) the applicant. This document also includes a qualitative summary of the application's strengths, weaknesses, and overall comments. Those applications that fall below "Promising" are encouraged to submit additional data. The identities of the reviewers *always* remain anonymous.

CONCLUDING REMARKS

Applicants may be interested in knowing that, across all applications to NREP, the relationship between utility and integrity has yielded a correlation of .78. The criteria most strongly associated with *utility* are attrition, general threats to validity, and issues pertaining to design; those most pertinent to *integrity* are design, threats to validity, and analysis. These correlations, which are displayed in the table below, suggest those criteria that discriminate most among applications and to which evaluators should pay close attention.

Correlations of Key NREP Criteria with Utility and Integrity

Key NREP Criteria	Utility	Integrity
Theory	0.34	0.47
Fidelity	0.47	0.54
Attrition	0.80	0.56
Design	0.66	0.91
Outcomes	0.40	0.53
Analysis	0.52	0.70
Threats	0.69	0.86

Applying for NREP status is a process. The NREP staff and SAMHSA extend their assistance to applicants in this process and look to you as future partners in sharing "what works" with the field.